CKME136 – Capstone Project Initial Results

Comparative Analysis for Walmart Sales Forecast

Adrian Wong (501030070)

2020-06-09

Github Link: https://github.com/adrianmwong/CKME136-Capstone-Project

Comparative Analysis for Walmart Sales Forecast

Introduction

Walmart Inc. is a multinational retail company that has more than ten-thousands stores worldwide. Selling all sorts of products including, grocery, household items, furniture, clothing, jewelry, electronics, and more. It is always hard to predict accurate sales for a huge company like Walmart, as many hidden factors could affect sales. Since the past centuries, various predictive methods and models have been developed, and Walmart Inc. must choose an accurate method for predicting future sales as it influences management decisions. With an accurate prediction of sales, management could better allocate resources for marketing and finding out which departments are doing poorly to maximize profit. Using Walmart Dataset from Kaggle as a case study, this project seeks to address the following question:

With the use of deep learning approaches in time series data, does the predict of sales become more accurate compared to traditional time series methods and regression methods?

By computing time series methods: Auto Regressor (AR), Auto-Regressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and Holt's Winter seasonal method. Then, regression methods including: Decision Tree, Random Forest. Also, with deep learning of neural network methods, including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). All of the mentioned above methods would be compared against Weighted Mean Absolute Error (WMAE) as requested in Kaggle. For calculation of algorithms, RStudio would be used. For data visualization and presentation, Weka and Tableau would be used. The best model would be used to compute future sales.

Literature Review

Businesses have been trying to utilizing math and different technique to predict sales, with accurate prediction, it is very helpful for them to prepare how much to order and for marketing. Throughout the centuries, many methods are being developed and used for predictions. It is always a hard choice to choose what methods as each technique has its strength and limitation (Mentzer & Moon, 2004; Lasek et al., 2016).

Timeseries and Hybrid Methods with Neural Network:

When there is uncertainty in predicting future values that are given with time stamp, the time-series approach seems to work the best (Goyall et al., 2018). It is not surprising that a researcher found out that more than 50% of the research papers on predictive analytics for the supply chain used the time series forecasting approach (Bonnes, 2014). By using traditional simple time series methods such as Autoregression (AR) and MA, a lot of the times the accuracy is not as desirable (Udom & Phumchusri, 2014). Researchers have stepped up and combining or modifying existed methods, such as developed SARIMA and SVR, which shows better results (Pai & Lin, 2005; Pai et al., 2010). In the recent decade, researchers utilized unsupervised machine learning and deep learning in time series modelings such as LSTM and GRU. Even though the main weakness of LSTM and GRU are their explainability, --as there is not much justification provided for these models (Baccar et al., 2019). Most researchers suggested that if the methods are trained and configurated properly, it yields better results than traditional approaches (Längkvist et al., 2014; Bandara et al., 2019).

Regressions and Hybrid Methods:

On the other hand, some researchers argue that sales prediction is rather a regression problem, this is because the time-series approach analyzes time series sequences and investigate what are the parameter statistics and other characteristics, and used that to predict future values using historical observed value. Whereas regression, it is more like testing theories, checking whether the current value in one or more independent time series influences the current value in another time series (Pavlyshenko, 2019). Similar to the time-series approach, the simpler regression models such as Ordinary Least Squares (OLS) is never the best methods for predicting sales. More sophisticated models like multiple linear regression, decision tree, random forest, and XGBoost outperform simple regression models (Wu et al.,2018; Hülsmann et al., 2012; Jain et al., 2015).

Comparison of both approaches:

Regardless of using the time-series approach or regression approach, one thing for sure is that there are never the “best” methods, as each dataset is different, and most of these methods and techniques have their strengths and weaknesses. Testing and comparing the results are needed to see which yields the best results. Therefore, in this research paper, the main focuses are to determine which methods work the best with the Walmart Sales forecast and would be ranked based on WMAE.

Dataset

In the Kaggle dataset, Walmart has provided historical weekly sales data extracted from the 45 Walmart store from 05/02/2010 to 26/10/2012, a total of 421,570 instances. Each of the stores contains around 90 departments. Additional to date and weekly sales amount, the dataset also included other variables such as store size, type, temperature, fuel price, consumer price index (CPI), unemployment date, and Walmart promotional markdown event date. The challenge for this paper is to predict department-wide sales for each store and how does promotional markdown events (Super Bowl, Labor Day, Thanksgiving, and Christmas) affect the prediction, as Walmart decided that the holiday weeks are weighted five times higher than non-holiday weeks.

Dataset provided in Kaggle: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

There are five CSV files included in the dataset: Features, Stores, Train, Test. and SampleSubmission (will not be used, as it is for Kaggle submission only). Total of 16 attributes noted, and description of the attributes listed below (Table 1):

Attributes:	Description:
Store	This is the store number, it ranged from 1-45
Dept	It is the department number, it ranged from 1-99, for different categories of items
Date	The Week
IsHoliday	Whether that specific week has special holiday in it
Weekly_Sales	Store weekly total amount in USD
Temperature	Average weekly temperature of the specific store region in Fahrenheit
Fuel_Price	Cost of Fuel of the specific store region in USD
MarkDown1-5	Anonymized Data related to the promotional markdown of Walmart, the data is only available after November 2011, and not all store has it
CPI	Consumer Price Index of specific store region for the week
Unemployment	Unemployment rate of specific store region for the month
Type	Type of the store, A, B or C
Size	Size of the specific store measured in square feet

Attributes and Descriptions: Table 1

Methodology

In this section, a general description of the methodology for this research paper would be discussed. Followed by the data analytics techniques that would be used and performance evaluation metrics that would apply to these data analytics techniques.

The goal of this research paper is to use historical data to predict the future sales of Walmart stores and departments using traditional time series approach, regression approach, and neural network approach.

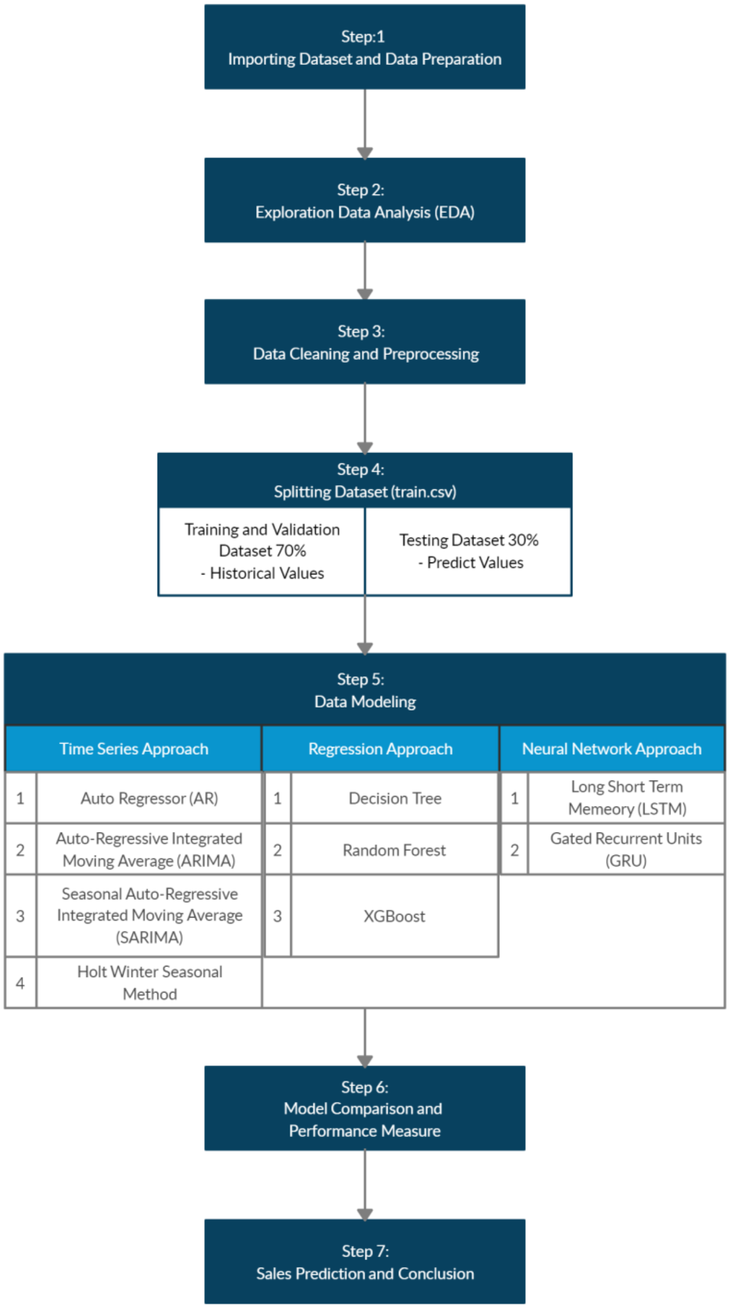

Step:1
Importing Dataset and Data Preparation
Step 2:
Exploration Data Analysis (EDA)
Step 3:
Data Cleaning and Preprocessing
Step 4:
Splitting Dataset (train.csv)
Training and Validation Dataset 70%
- Historical Values
Testing Dataset 30%
- Predict Values
Step 5:
Data Modeling
Time Series Approach
Regression Approach
Neural Network Approach
1
Auto Regressor (AR)
2
Auto-Regressive Integrated Moving Average (ARIMA)
3
Seasonal Auto-Regressive Integrated Moving Average (SARIMA)
4
Holt Winter Seasonal Method
1
Decision Tree
2
Random Forest
3
XGBoost
1
Long Short Term Memeory (LSTM)
2
Gated Recurrent Units (GRU)
Step 6:
Model Comparison and Performance Measure
Step 7:
Sales Prediction and Conclusion

Fig.1 Schematic representation of methodology

The methodology approach is illustrated in Fig.1. In Step 1: Data import and data preparation. The csv files would be loaded the dataset into Jupyter notebook, then it would be reviewed for understanding purpose only, such as how many variables are in it and number of instances. The relevant csv files would be merged with the inner join.

Step 2: Exploratory Data Analysis (EDA). After merging of the csv file, EDA would be performed. Descriptive statics would be used to investigate the count, mean, minimum, maximum, interquartile ranges, and number of NAs. The paper would also investigate the relations between sales (dependent variable) and other independent variables with visual aids.

Step 3: Data cleaning and preprocessing. Based on the previous analysis of dependent variables and independent variables. The data would be clean, including replacing NAs, Boolean and category type data to integers, and set Date as Index. Then, a correlation heat map would be used to see the correlations of all the attributes. In the end, low correlations attributes would be removed from the dataset.

Step 4: Splitting dataset. The CSV file included all historical data would be split into two part in a ration of 70:30 for the training set and testing set respectively.

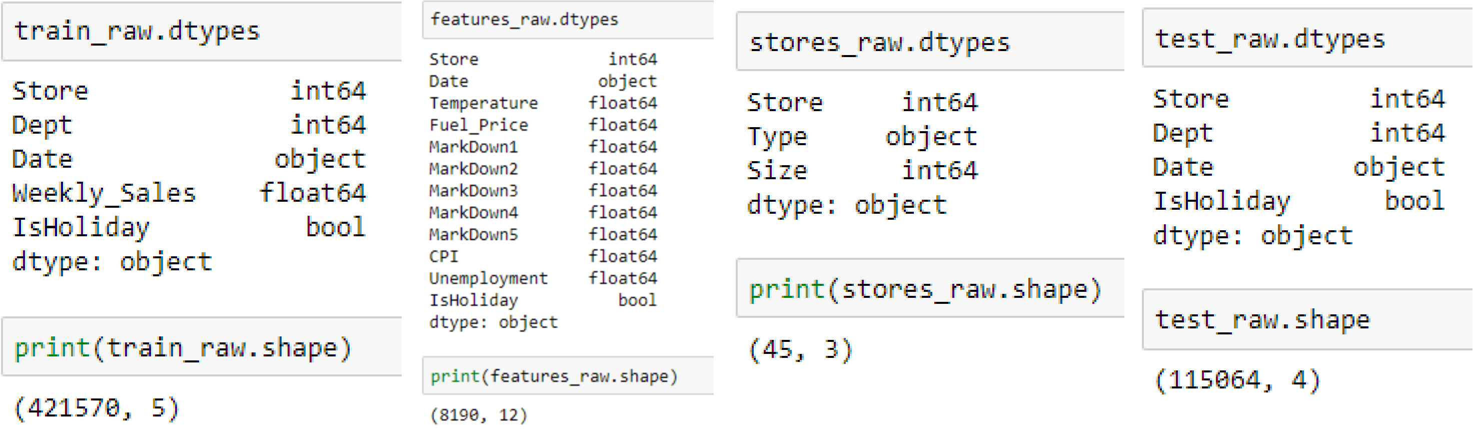
Step 5: Data modeling. The training set would be fitted in all the suggested models in time series approaches, regression approaches, and neural network approaches. In the time-series approach, common methods included AR, ARIMA, SARIMA, and Holt Winter Seasonal Method would be used. In the Regression approach, common methods included Decision Tree, Random Forest Regressor, and XGBoost would be used. Last but not least, the neural network approach including LSTM and GRU would be used.

Step 6: Model comparison and performance measure. After each of the methods listed model developed. All of the models would be used to compare against each other with performance metrics, such as Mean Absolute Error (MAE) and Mean Squared Error (MSE).

Step 7: Sales Prediction and conclusion. Based on the previous step, the lowest MAE or MSE scored would be used to predict future weekly sales. Recommendation and limitation would be used to conclude the research paper.

Import Dataset and Data Preparation

Since the Dataset section above already investigated the nature of the attributes, here would just show where they located in each CSV file. Below are the four CSV files including their attributes and number of instances. The train CSV file is the main CSV file that would be used for data modeling and testing. The feature and store CSV file are additional features that Walmart included that may affect the weekly sales data, therefore would be merged to train CSV files later on. The test CSV is the actual file used to predict sales, so a total of 115,064 instances would be predicted in a later section.



```
train_raw.dtypes

Store             int64
Dept              int64
Date             object
Weekly_Sales    float64
IsHoliday          bool
dtype: object

print(train_raw.shape)

(421570, 5)
```

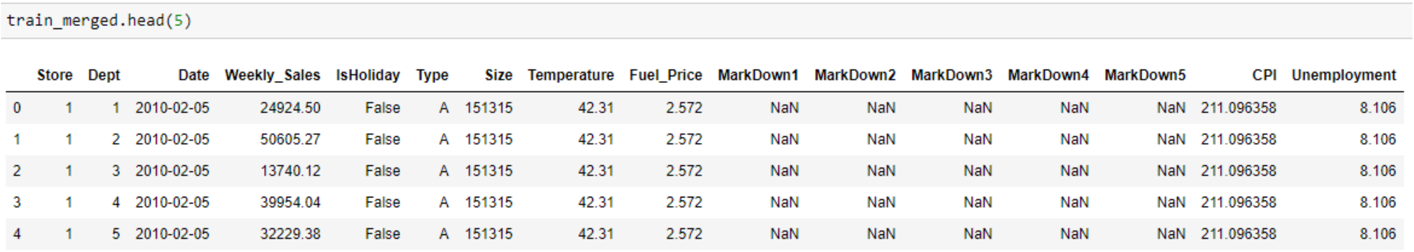
```
features_raw.dtypes

Store             int64
Date             object
Temperature     float64
Fuel_Price      float64
MarkDown1       float64
MarkDown2       float64
MarkDown3       float64
MarkDown4       float64
MarkDown5       float64
CPI             float64
Unemployment    float64
IsHoliday          bool
dtype: object

print(features_raw.shape)

(8190, 12)
```

```
stores_raw.dtypes

Store     int64
Type     object
Size      int64
dtype: object

print(stores_raw.shape)

(45, 3)
```

```
test_raw.dtypes

Store         int64
Dept          int64
Date         object
IsHoliday      bool
dtype: object

test_raw.shape

(115064, 4)
```

Fig.2 Shape of the four CSV files

The method used for merging train.csv, feature.csv, and stores.csv are inner join based on the attribute of "Store", "Date", and "IsHoliday". Below are the first five instanes for preview. Total of 421,570 instances 16 attributes.



```
train_merged.head(5)
```

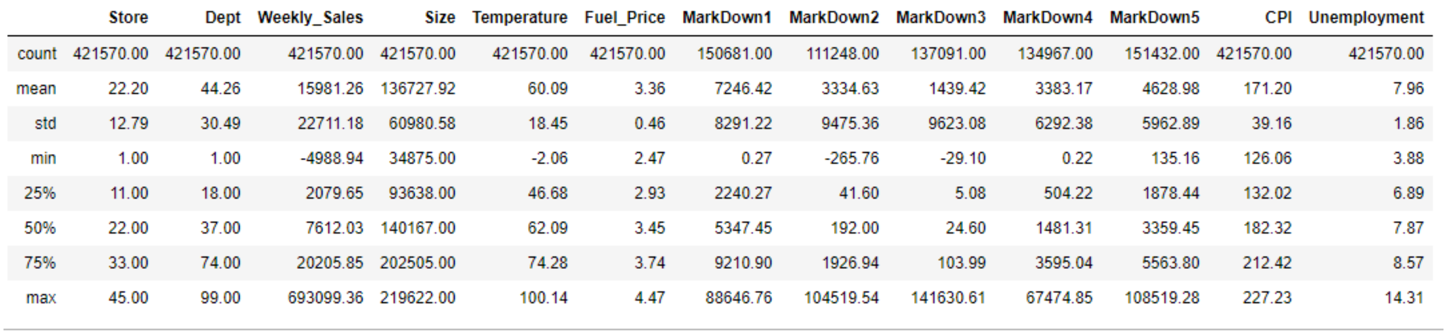
	Store	Dept	Date	Weekly_Sales	IsHoliday	Type	Size	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment
0	1	1	2010-02-05	24924.50	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
1	1	2	2010-02-05	50605.27	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
2	1	3	2010-02-05	13740.12	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
3	1	4	2010-02-05	39954.04	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106
4	1	5	2010-02-05	32229.38	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106

Fig.3 First 5 row of merged train CSV file

Exploratory Data Analysis (EDA)

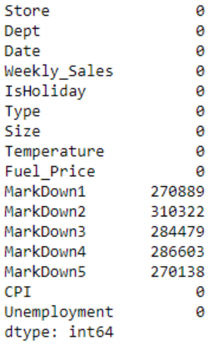
Descriptive Statistics:

Based on the descriptive statistic chat generated below (Fig.4), most of the numeric attribute looks reasonable. Store number, department number, size, temperature, fuel price, CPI and unemployment rate does not have negative number and does not have missing data. Whereas for Markdown 1 to 5, more than 200k data are missing, and this is because Markdown 1 to 5 would only be recorded when there are promotional events happen in certain store of Walmart (Fig.5).



	Store	Dept	Weekly_Sales	Size	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment
count	421570.00	421570.00	421570.00	421570.00	421570.00	421570.00	150681.00	111248.00	137091.00	134967.00	151432.00	421570.00	421570.00
mean	22.20	44.26	15981.26	136727.92	60.09	3.36	7246.42	3334.63	1439.42	3383.17	4628.98	171.20	7.96
std	12.79	30.49	22711.18	60980.58	18.45	0.46	8291.22	9475.36	9623.08	6292.38	5962.89	39.16	1.86
min	1.00	1.00	-4988.94	34875.00	-2.06	2.47	0.27	-265.76	-29.10	0.22	135.16	126.06	3.88
25%	11.00	18.00	2079.65	93638.00	46.68	2.93	2240.27	41.60	5.08	504.22	1878.44	132.02	6.89
50%	22.00	37.00	7612.03	140167.00	62.09	3.45	5347.45	192.00	24.60	1481.31	3359.45	182.32	7.87
75%	33.00	74.00	20205.85	202505.00	74.28	3.74	9210.90	1926.94	103.99	3595.04	5563.80	212.42	8.57
max	45.00	99.00	693099.36	219622.00	100.14	4.47	88646.76	104519.54	141630.61	67474.85	108519.28	227.23	14.31

Table 4 Descriptive statistic for the numeric attributes



```
Store                0
Dept                 0
Date                 0
Weekly_Sales         0
IsHoliday            0
Type                 0
Size                 0
Temperature          0
Fuel_Price           0
MarkDown1       270889
MarkDown2       310322
MarkDown3       284479
MarkDown4       286603
MarkDown5       270138
CPI                  0
Unemployment         0
dtype: int64
```

Fig.5 NAs in the merged dataset

Outliers:

By running boxplot (Fig. 6), it is noticeable that there are 7 attributes that have many outliers, which included Weekly_Sales, Markdown 1 to 5 and unemployment rate. No outliers has been removed as it may contain valuable information for the analysis.

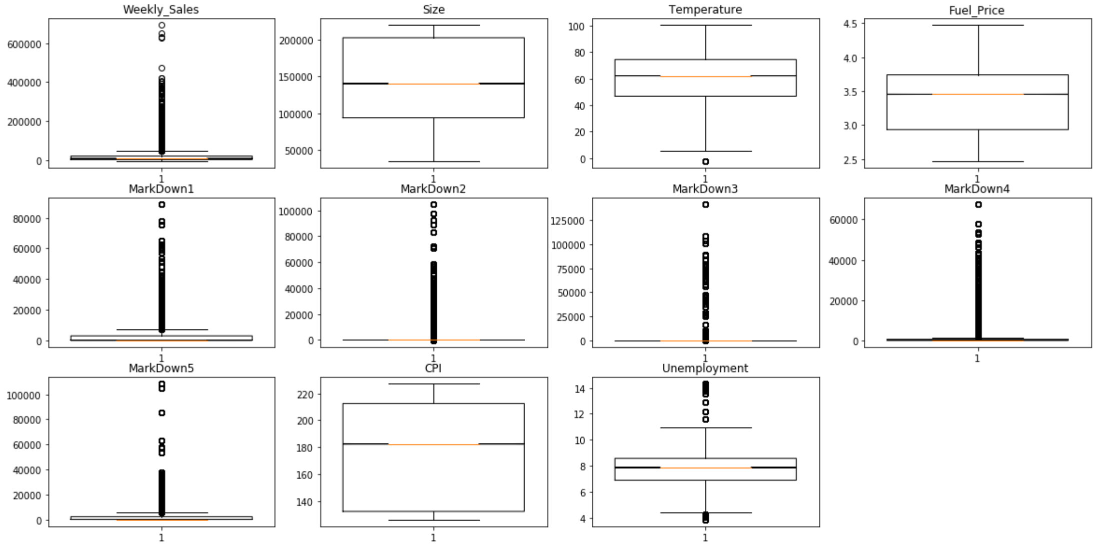


Fig.6 Outliers

Top and Bottom Five Stores in Average Sales:

By calculating the average weekly sales of each store, the highest contributors are store number 20, 4, 14, 13, and 2, each of these stores are able to generate more than $25,000 each week during 2010 to 2012 (Fig.7a). Those five stores total sales accounted for more than 22% of the total sales for 2010 to 2012 (Table 7b). Whereas store number 5, 33, 44, 3, and 38 only contributed $5,000 to $7,000 each week only (Fig. 7a), adding all those bottoms five stores only contributed around 4% of the total sales.

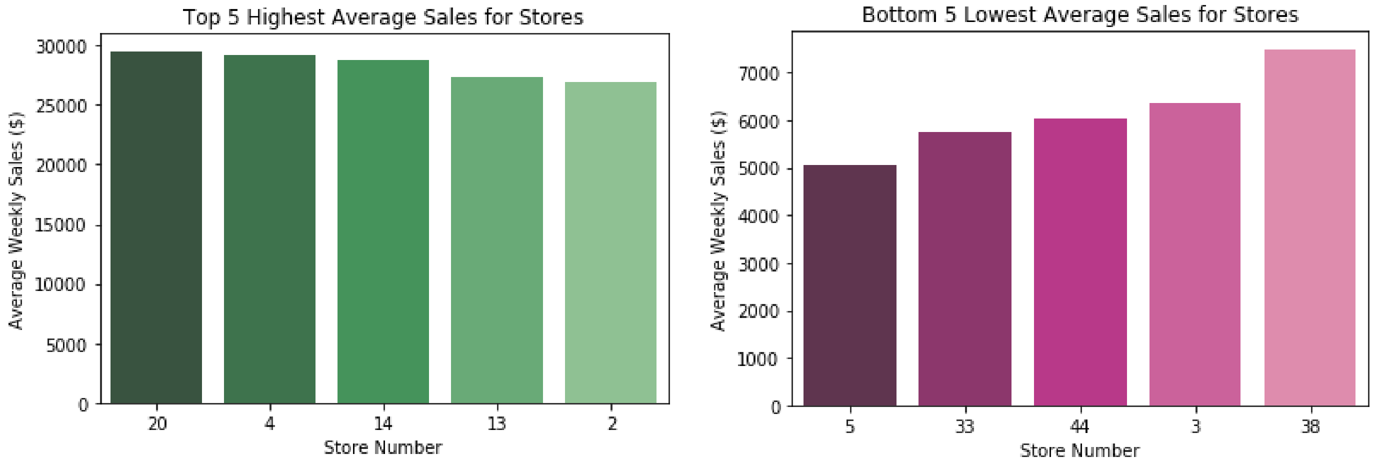


Fig.7a Top and bottom five stores in sales

Top Store	Store Total Sales	Proportion to Total sales	Top Store	Store Total Sales	Proportion to Total sales
20	301,397,792	4.47%	5	45,475,689	0.67%
4	299,543,953	4.45%	33	37,160,222	0.55%
14	288,999,911	4.29%	44	43,293,088	0.64%
13	286,517,704	4.25%	3	57,586,735	0.85%
2	275,382,441	4.09%	38	55,159,626	0.82%
Sum	1,451,841,802	22%	Sum	238,675,360	4%
Total Sales of 2010-2012:		6,737,218,987			

Table.7b Store total sales contribution

Top and Bottom Five Departments in Average Sales:

In the perspective of sales in departments, the top five departments are 92, 95, 38, 72, and 65, making average sales of more than $40,000 each week (Fig. 8a). To investigate further, based on a number list provided by Walmart[1], the department names for the top five deparments for 92, 95, 38, 72, and 65 are Grocery, DSD Grocery, Pharmacy RX, Electronics, and Gasoline respectively. Whereas, bottom five departments are 51, 39, 78, 43, 47, each with less than $25 a week. Out of the five departments, only department 39 is on the list and the name is customer service, the rest would be special departments which is not commonly used. Overall, it is reasonable that the respective departments are the highest sales and lowest sales based on the nature.

[1] https://blog.8thandwalton.com/wp-content/uploads/2017/10/Walmart-Department-Numbers.pdf?utm_campaign=Lead%20Generation&utm_medium=email&_hsenc=p2ANqtz--CCzb0keOFMueCFNOrvAAwqcC6gIgcCKebbFQlwPKWD6uCd4srCUwVXlUy64gBJpKUrxKF1DWppwa_q2yszjHxto7NZQ&_hsmi=59683228&utm_content=59683228&utm_source=hs_automation&hsCtaTracking=cdf6bba2-a9f4-4034-8411-9d510d02dc99%7C04f3a820-62fb-4901-9bb1-3b3d0f8161ec

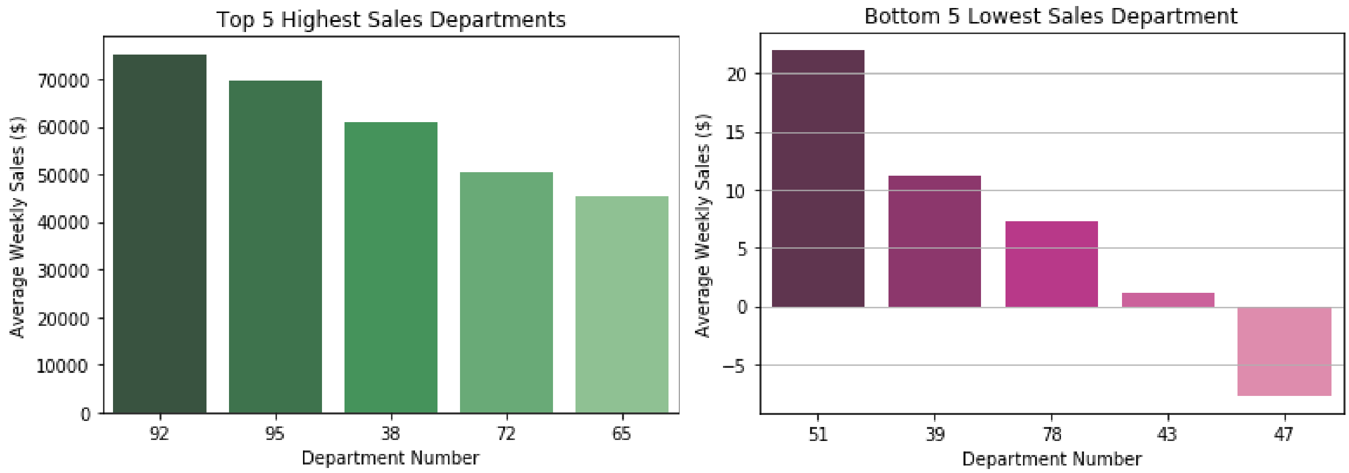


Fig.8 Top and Bottom Five Departments in Average Sales

Comparison of Stores and Departments in Average Weekly Sales

By creating a heat map between stores and departments, it would be easier to visualize the relationships and respective average weekly sales performance (Fig. 9). Departments of 38, 40 and 88 to 95 generated the most sales.

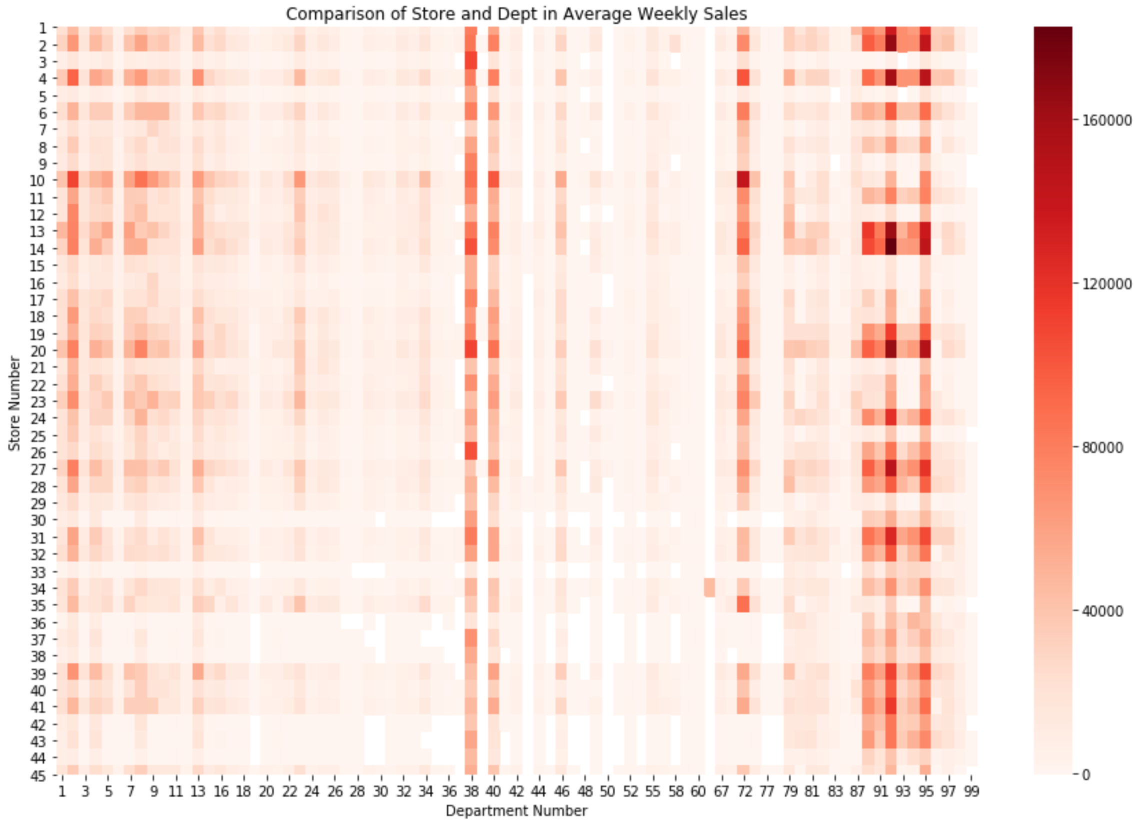


Fig.9 Comparison of stores and departments in average weekly sales

Weekly Sales Comparison for the Dataset (2010,2011,2012)

From the below figure, all three years weekly sales are very similar. Especially when there are holiday with Super Bowl, Labor Day, and Thanksgiving, the sales increases dramatically with an exception of Christmas. The reason is because people would buy gift for boxing day which is prior to the day after Christmas Day (holiday). Also, it is noticible that Walmart did not provide other holidays, such Easter Holiday for week 13 or 14, which also have high sale.

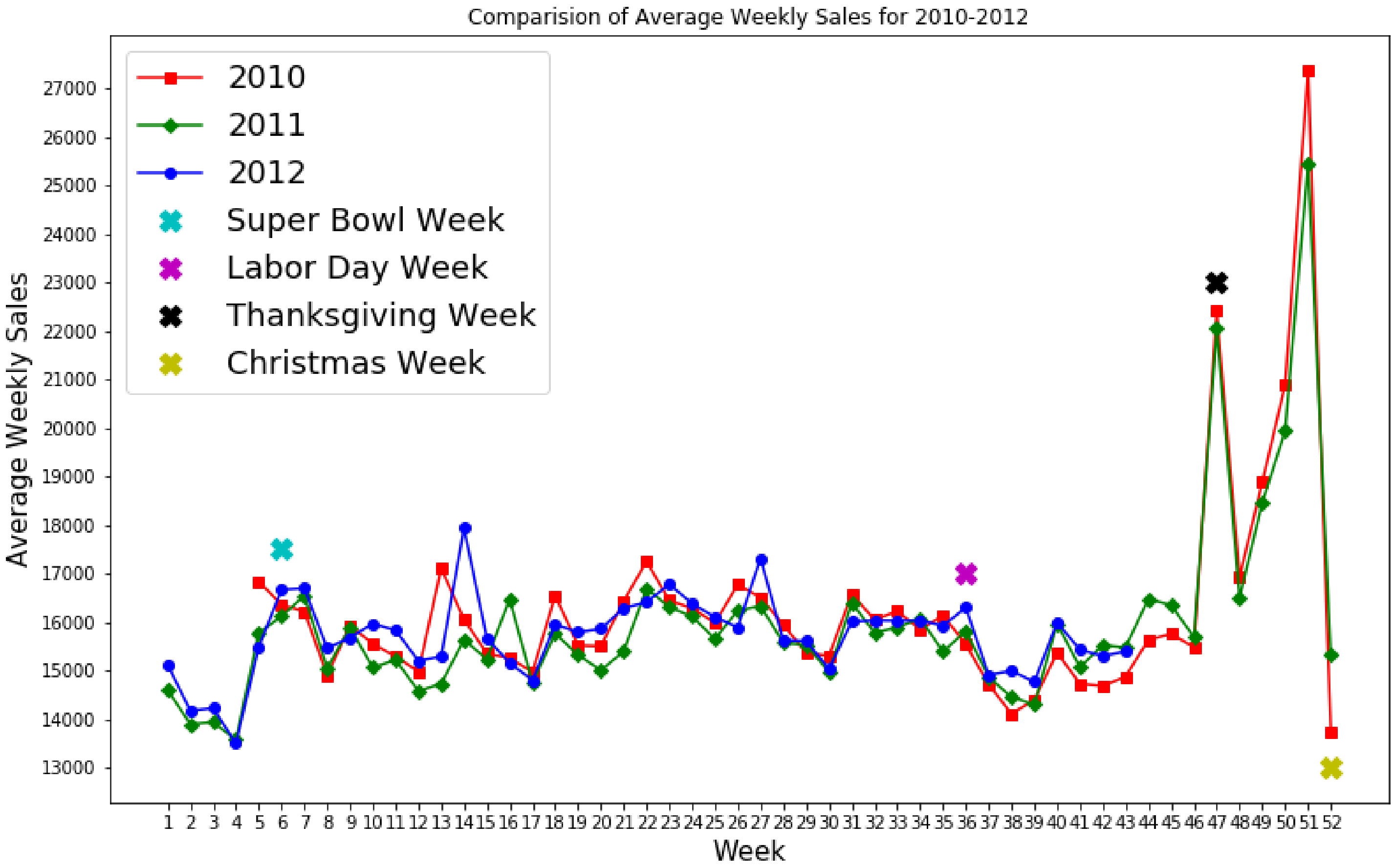


Fig. 10 Comparison of average weekly sales for 2010-2012

Comparing Attribute of IsHoliday and Weekly_Sales:

When comparing the attribute of IsHoliday and Weekly_Sales, based on the figure below (Fig. 11), weeks that has holiday are indeed generating higher sales. Having a closer look at the box plot, although seems insignificant, week that are consider as holiday still have slighter higher sales. Also based on Fig. 10, week number 52 is a special week, for better prediction, isntead of marking week 52 is holiday week, week 51 should be marked as holiday.

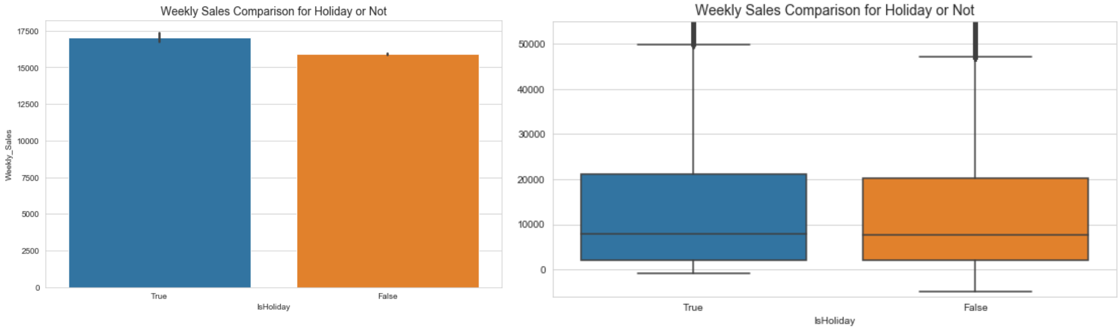
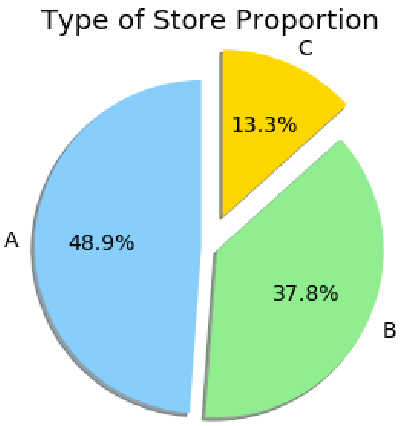


Fig.11 Weekly comparison for IsHoliday or not

Comparing the Attribute Type and Store:

Walmart did not provide information what the attribute “Type” is, but for now, based on the total 45 stores, around 49% of stores are classified as type A; 38% classified as type B; and 13% for type C (Fig.12a). From Fig.12b, it is clear that stores that are labeled as type “A” generated the highest average weekly sales, based on the mean. Type “B” generated lesser sales compared to “A”, but higher than “C”. To conclude, the order of highest weekly sales for types are A>B>C.



```
      Store
Type
A        22
B        17
C         6
```

Fig.12a Type of store proportion

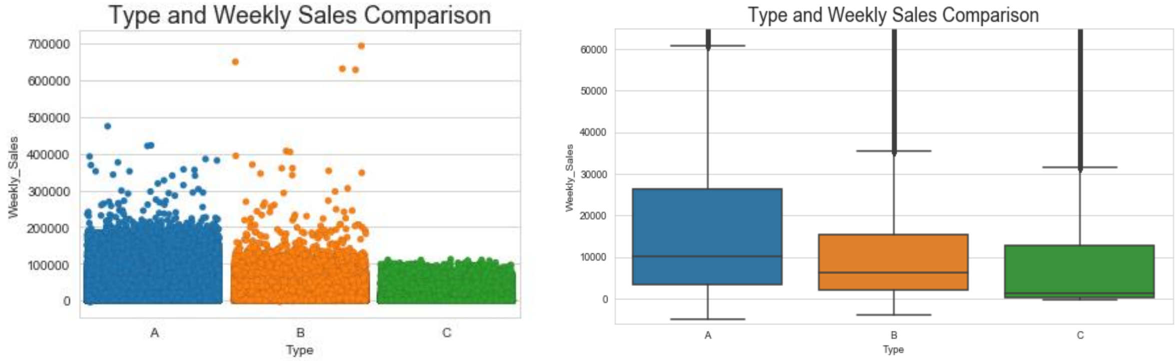


Fig.12b Type and Weekly Sales Comparison

Type and Size Comparison:

Walmart may also label store based on their size. Based on the below figure, disregard to some of the outliers, it is clear that type “A” store is the biggest, with a mean of more than 200,000 squared feet. Then type “B”, with average of 110,000 squared feet. Lastly, type “C”, that are less than 50,000 squared feet.

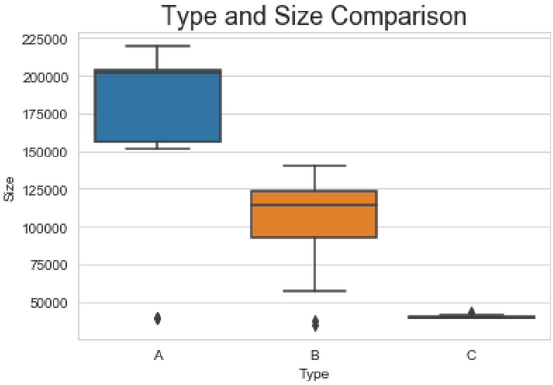


Fig.13 Type and size comparison

Data Cleaning and Preprocess

Based on previous analysis on the dependent variables and independent variables, the following section would be cleaning the data including following procedures.

Adding New Attribute of “Year” and “Week”

By adding the attribute of “Year” and “Week”, it would be easier to separate the data base on year or week and for visualization.

Replacing “Type” outputs of A, B, C to 3, 2, 1

In the previous section, store label A indicate higher sales than B, then C. Therefore, higher sales stored that are labeled as A would be replaced with 3, B would be replaced as 2, and C would be replaced as 1 for easier comparison in correlations heatmap in the later section.

Replacing “IsHoliday” from Boolean Values to Numeric

By replacing the value of true and false, the data would be easier to model and present.

Replacing value of “IsHoliday” Week 51 and Week 52 to 1 and 0 Respectively

As in the previous section mentioned, week 51 and 52 “IsHoliday” value would be replaced to 1 and 0 as the sales are caused by Christmas but people purchase the gift before the holiday week.

Making the "Date" as Index

For easier management for splitting the dataset into train and test set for time series.

Visualizing the Correlations of the Attributes and Removing Attributes

In the figure below, all of the attributes are compared against the “Weekly_Sales” attribute as this is the dependent variable. Independent variables such as “Temperature”, “Fuel_Price”, “CPI” and “Unemployment” have very low correlation with “Weekly_Sales”, therefore it would be removed. Markdown 1-5 would also be removed as there are many missing values and low correlation as well. Even though the variable of “IsHoliday” have low correlations to weekly sales, it would not be removed as it is an important attribute for classifying which week have holiday and is used for WMAE performance metrics. Same for “Store”, “Dept”, “Year” and “Week”, those attributes are important for separating the data set and would not be removed. Lastly, the attribute of “Type” and “Size” have a weak positive linear relationship to Weekly Sales, therefore would not be removed.

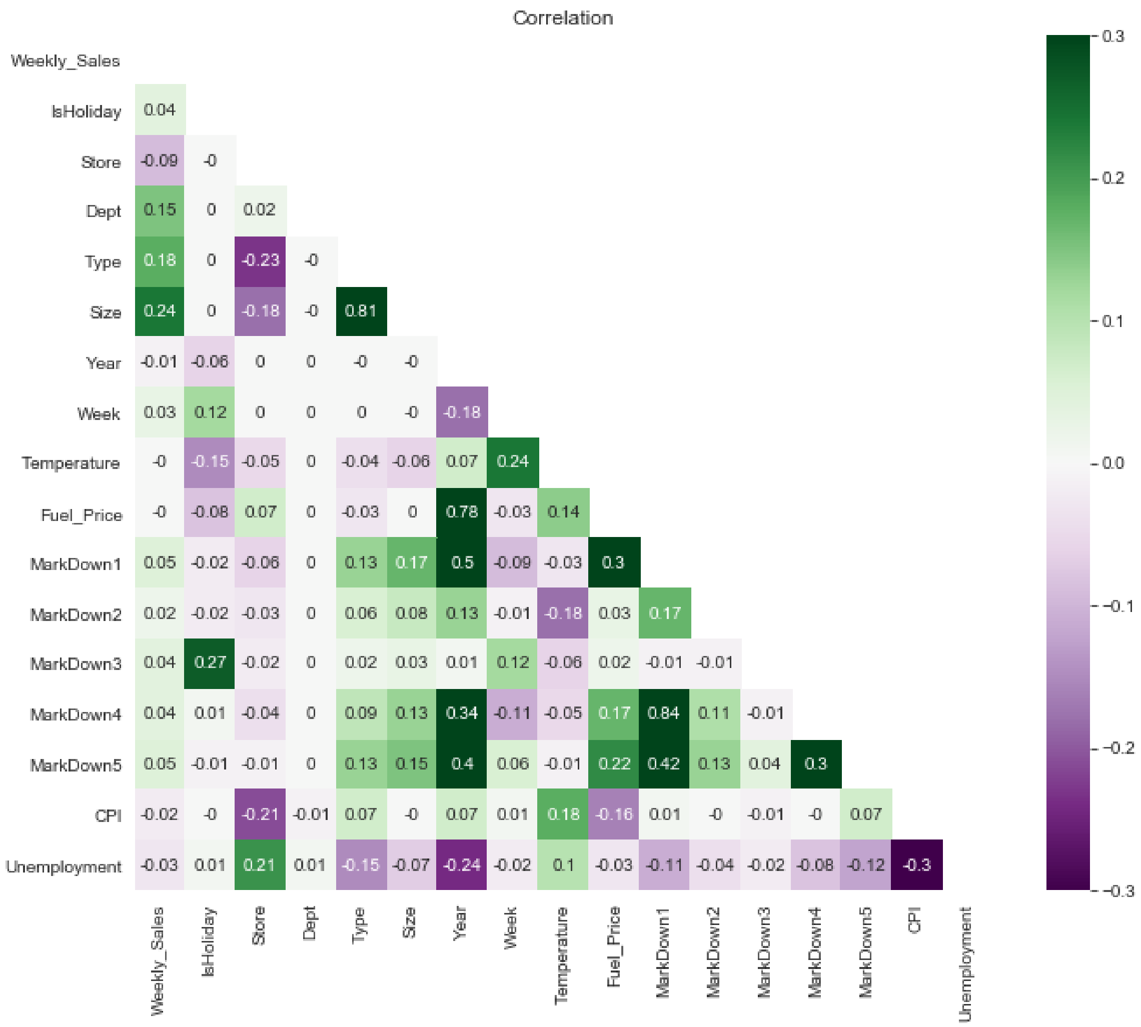
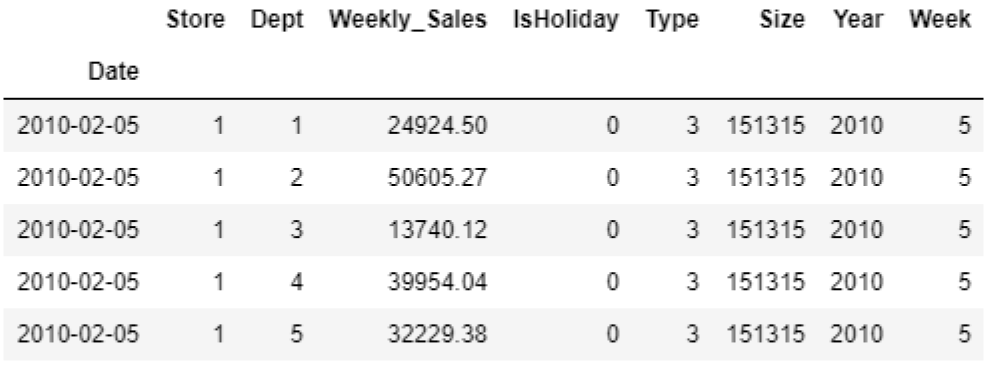


Fig.14 Correlation heatmap

Final Output for Modeling:

Below is the preview of the final output CSV for modeling, only 8 attributes remains and all are numerical values.



	Store	Dept	Weekly_Sales	IsHoliday	Type	Size	Year	Week
Date								
2010-02-05	1	1	24924.50	0	3	151315	2010	5
2010-02-05	1	2	50605.27	0	3	151315	2010	5
2010-02-05	1	3	13740.12	0	3	151315	2010	5
2010-02-05	1	4	39954.04	0	3	151315	2010	5
2010-02-05	1	5	32229.38	0	3	151315	2010	5

Fig.15 Final output CSV for modeling

Splitting Dataset

The dataset from the “train.csv” would be spitted into train: test and train: cross validate sets in a ratio of 70:30 and would be arranged based on the date as shown below.

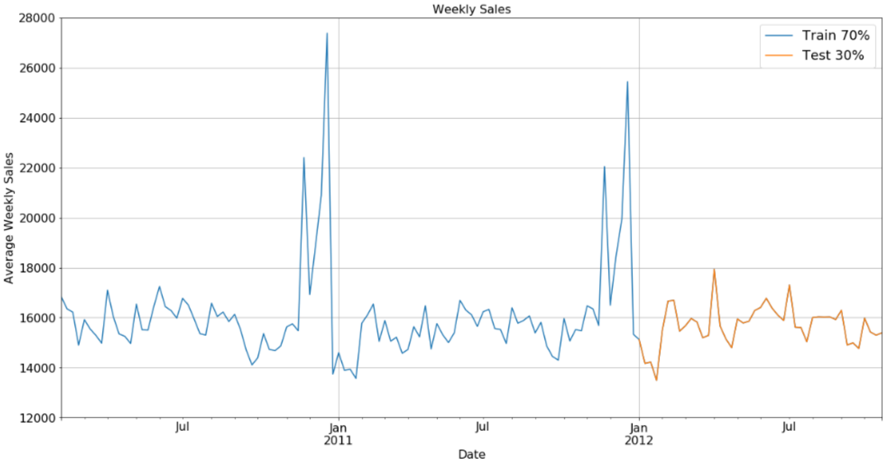


Fig.16 Split Training and Test set

Data Modeling – Time Series Approach:

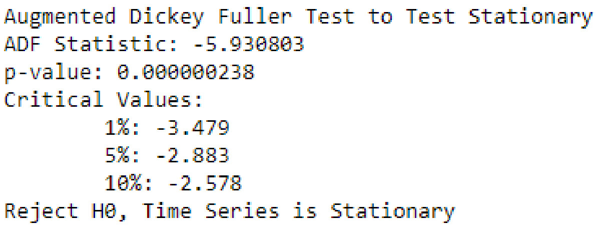
Since the main purpose of the research paper is to predict further sales for specific date, store, and departments. Also, it has provided the data in a time series format, therefore popular time series forecasting methods would be implemented, including AR, ARIMA, SARIMA and Holt Winter Seasonal Method.

In time series model, it assumes that the data are stationary, meaning that it has constant mean, variance and covariance. If the dataset is not stationary, adjustment such as differentiation or transformation are needed.

Augmented Dickey-Fuller Test (ADF):

To test whether the data are stationary, ADF test are used to test stationarity. ADF is a method used to test whether the data has a unit root present in the time series data, it also included the lagged terms for determination. It tests the dataset using hypothesis. H0 suggest unit root is presented, whereas H1 suggest that the data is stationary. After running the ADF, it will provide an ADF statistics and a p-value. The value of the ADF statistic should be negative, and the more negative the number is, the higher the chance of rejecting H0.

Below figure shows the ADF statics and p-value after running. The ADF statistic are -5.93 and the p-value is near zero. For the purpose of the research paper, 95% confidence level are used. Since the p-value is lesser than 5%, we rejected H0, meaning that the time series is stationary, and could be applied to AR, ARIMA, SARIMA model.



```
Augmented Dickey Fuller Test to Test Stationary
ADF Statistic: -5.930803
p-value: 0.000000238
Critical Values:
        1%: -3.479
        5%: -2.883
        10%: -2.578
Reject H0, Time Series is Stationary
```

Fig.17 ADF test

Auto Regressor (AR):

The data would first be predicted with AR, which is similar to linear regression model and predict value based on a linear combination of input values. Even though this method is simple and would not be a good predictor for the sales prediction, it is still worth looking before implementing the whole ARIMA and SARIMA.

In the below figure, after implementing the AR model from the train set to the test set, the prediction is shown in the red line. Further investigation by using performance metric would be in the comparison session.

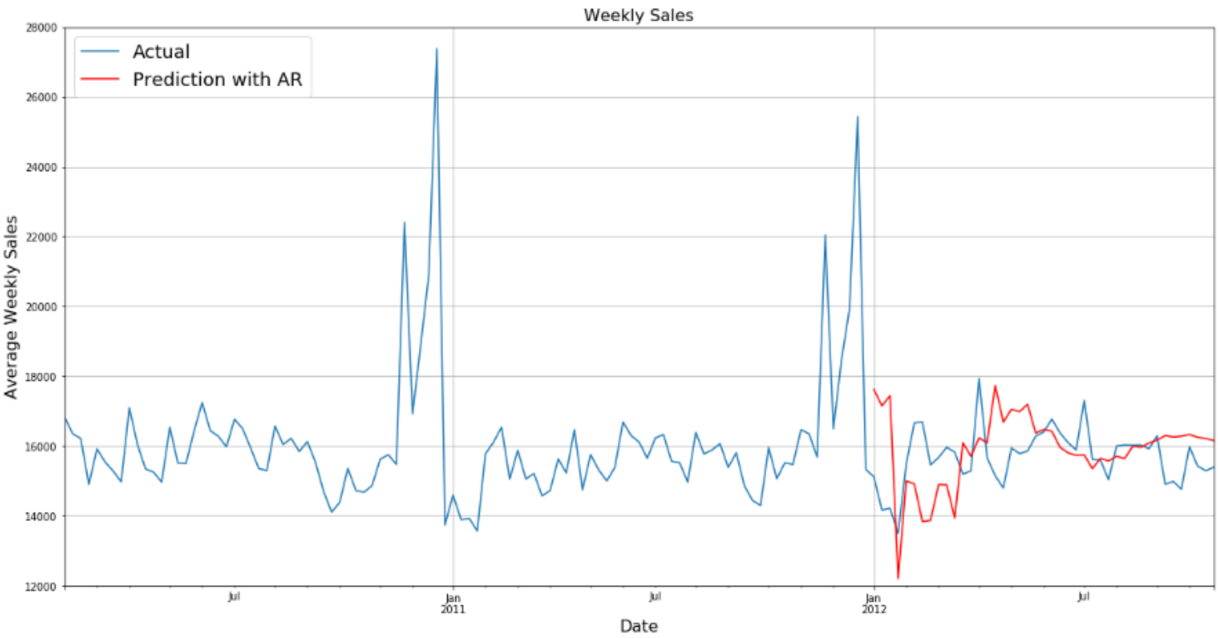


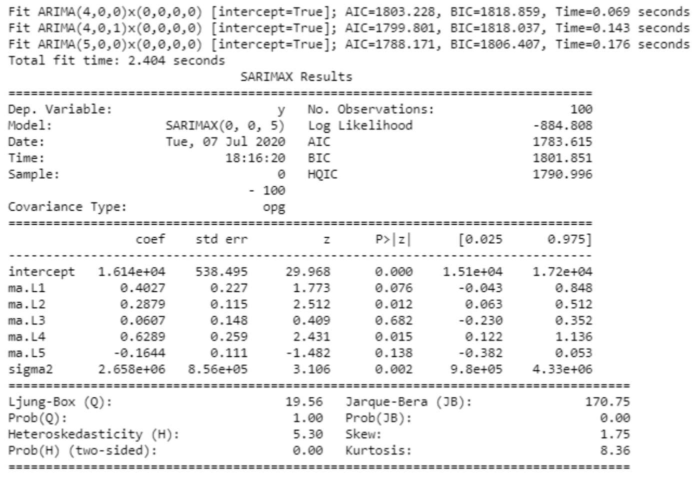
Fig.18 AR prediction compared to the actual data

Auto Regressive Integrated Moving Average (ARIMA):

ARIMA is similar to AR with more parameters, the most important part is to find the appropriate values for “p,d,q”. The value of p is the number of Auto Regressors (AR), where d is the difference (I), and q is the number of Moving Average (MA).

To find most appropriate number for the “p, d, q”, one method is to plot Auto Correlated Function (PCF) and Partial Auto Correlated Function (PACF) graph and pick the numbers. Another method, which is used in this research paper, is to create a function that test the combination values of “p, d, q” and suggest the best combination based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

The figure in Fig.19 show the final result of the function used to select the best combination of “p, d, q”, which is 0, 0, 5 respectively, as it has the lowest average AIC and BIC value. Fig.20 show the diagnostics of the current parameters with residual plots. The Top left graph is showing that the residual errors seems fluctuate near zero except for sales near the year. The top right suggests the residuals are almost normal distribution with a mean near zero. The bottom left shows that the distribution is skewed as some data are way off the red line. The bottom right ACF plot shows that the residual errors are not autocorrelated, this current model are seeming able to explain the all pattern. Fig.21 show how the combination of p=0, d=0, q=5 and how the prediction looks compared to the actual dataset. Comparing to the AR model earlier, the ARIMA provided a straight line, which would not able to predict yearend sales, but further investigation by using performance metric would be in the comparison session.



```
Fit ARIMA(4,0,0)x(0,0,0,0) [intercept=True]; AIC=1803.228, BIC=1818.859, Time=0.069 seconds
Fit ARIMA(4,0,1)x(0,0,0,0) [intercept=True]; AIC=1799.801, BIC=1818.037, Time=0.143 seconds
Fit ARIMA(5,0,0)x(0,0,0,0) [intercept=True]; AIC=1788.171, BIC=1806.407, Time=0.176 seconds
Total fit time: 2.404 seconds
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  100
Model:               SARIMAX(0, 0, 5)   Log Likelihood                -884.808
Date:                Tue, 07 Jul 2020   AIC                           1783.615
Time:                        18:16:20   BIC                           1801.851
Sample:                             0   HQIC                          1790.996
                                - 100
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept   1.614e+04    538.495     29.968      0.000    1.51e+04    1.72e+04
ma.L1          0.4027      0.227      1.773      0.076      -0.043       0.848
ma.L2          0.2879      0.115      2.512      0.012       0.063       0.512
ma.L3          0.0607      0.148      0.409      0.682      -0.230       0.352
ma.L4          0.6289      0.259      2.431      0.015       0.122       1.136
ma.L5         -0.1644      0.111     -1.482      0.138      -0.382       0.053
sigma2      2.658e+06   8.56e+05      3.106      0.002     9.8e+05    4.33e+06
===================================================================================
Ljung-Box (Q):                       19.56   Jarque-Bera (JB):               170.75
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               5.30   Skew:                             1.75
Prob(H) (two-sided):                  0.00   Kurtosis:                         8.36
===================================================================================
```

Fig.19 auto selector result

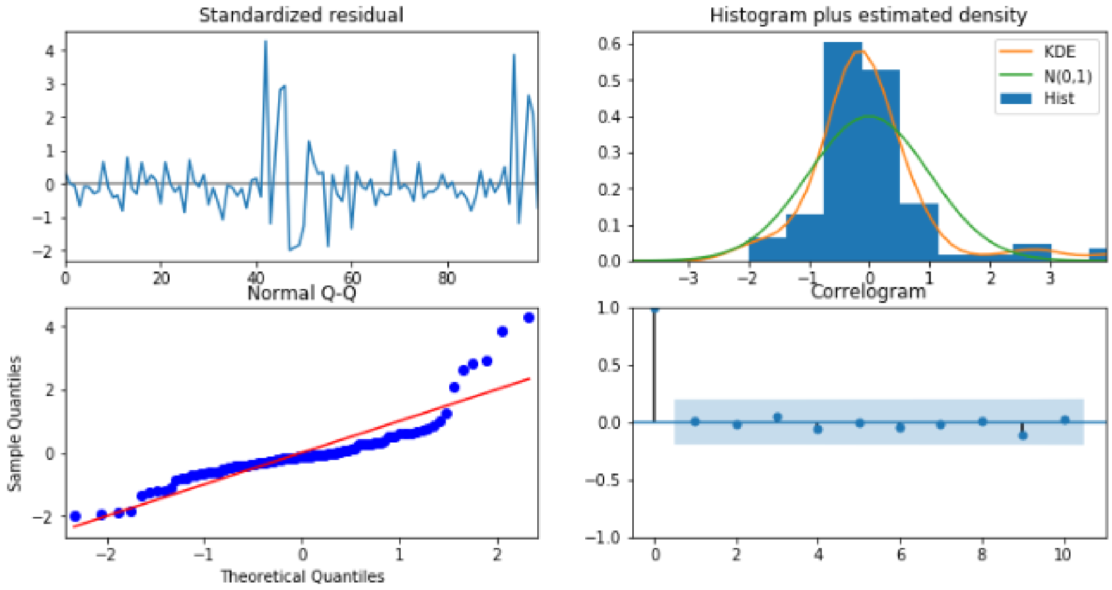


Fig.20 diagnostics of the ARIMA (0, 0, 5)

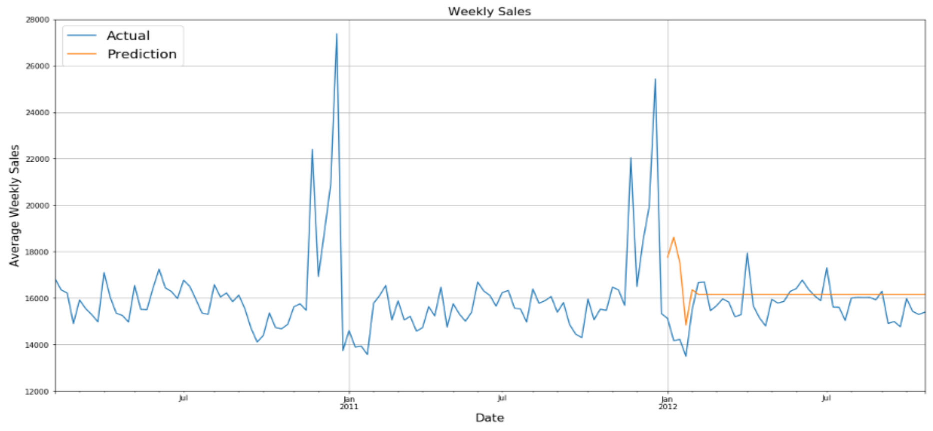


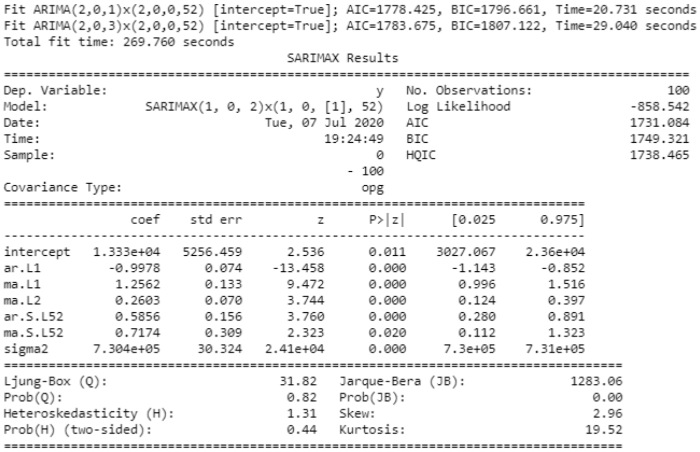
Fig.21 ARIMA compared to actual data

Seasonal Auto Regressive Integrated Moving Average (SARIMA):

Again, similar to AR and ARIMA, the SARIMA model included the seasonal element, which would further increase accuracy in this case. Additional to the parameters of p, d, and q, there are four more elements for the SARIMA, which are “P”, “D”, “Q”, and “s”. The “P” is number of autoregressive terms or lags for the time series. The “D,” is the differencing for stationary time series data. The “Q” is the number of moving average terms or lag of forecast errors. Lastly, “s” is the seasonal length for the data. In figure 22, similar to ARIMA, a grid search for the best combination for the data, which is SARIMA (1, 0, 2) (1, 0, 1, 52) as it has the lowest average of AIC and BIC based on the grid search.

In the residuals plots, the results are similar to ARIMA (0, 0, 5), but the standardized residual plot shows that the residuals are closer to zero compared to ARIMA. Also, the histogram and density graph are slightly flatter than ARIMA. The normal Q-Q graph has more residuals closer to the red line compared to the ARIMA. Overall, an improvement compared to ARIMA.

In figure 24, the predicted compared to the actual value looks more promising than the ARIMA, as it the predicted values are close to the actual value, unlike ARIMA, which is only a straight line. Further investigation by using performance metric would be in the later session.



```
Fit ARIMA(2,0,1)x(2,0,0,52) [intercept=True]; AIC=1778.425, BIC=1796.661, Time=20.731 seconds
Fit ARIMA(2,0,3)x(2,0,0,52) [intercept=True]; AIC=1783.675, BIC=1807.122, Time=29.040 seconds
Total fit time: 269.760 seconds
                                       SARIMAX Results
=============================================================================================
Dep. Variable:                                     y   No. Observations:                  100
Model:              SARIMAX(1, 0, 2)x(1, 0, [1], 52)   Log Likelihood                -858.542
Date:                               Tue, 07 Jul 2020   AIC                           1731.084
Time:                                       19:24:49   BIC                           1749.321
Sample:                                            0   HQIC                          1738.465
                                               - 100
Covariance Type:                                 opg
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
intercept    1.333e+04   5256.459      2.536      0.011    3027.067    2.36e+04
ar.L1          -0.9978      0.074    -13.458      0.000      -1.143      -0.852
ma.L1           1.2562      0.133      9.472      0.000       0.996       1.516
ma.L2           0.2603      0.070      3.744      0.000       0.124       0.397
ar.S.L52        0.5856      0.156      3.760      0.000       0.280       0.891
ma.S.L52        0.7174      0.309      2.323      0.020       0.112       1.323
sigma2       7.304e+05     30.324   2.41e+04      0.000     7.3e+05    7.31e+05
====================================================================================
Ljung-Box (Q):                        31.82   Jarque-Bera (JB):              1283.06
Prob(Q):                               0.82   Prob(JB):                         0.00
Heteroskedasticity (H):                1.31   Skew:                             2.96
Prob(H) (two-sided):                   0.44   Kurtosis:                        19.52
====================================================================================
```

Fig.22 auto selector result

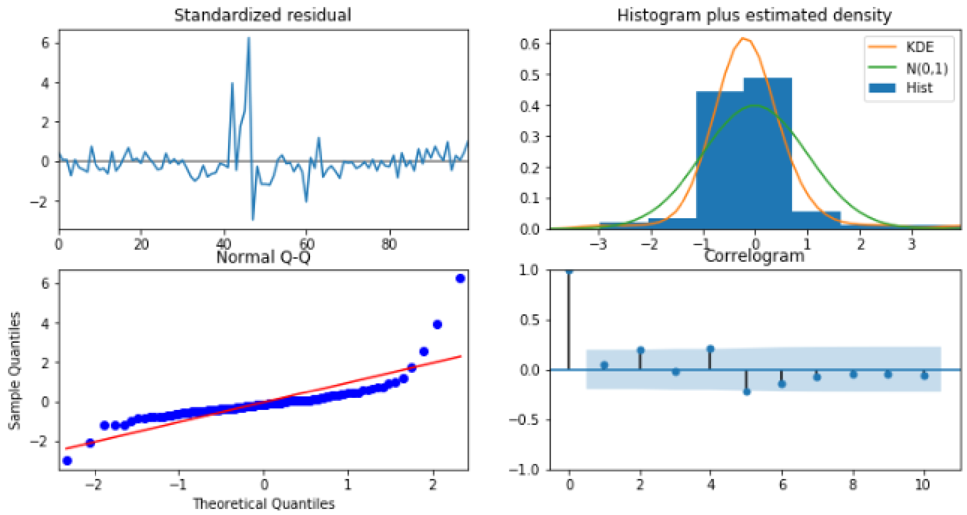


Fig.23 diagnostics of the SARIMA (1, 0, 2)x(1, 0, 1, 52)

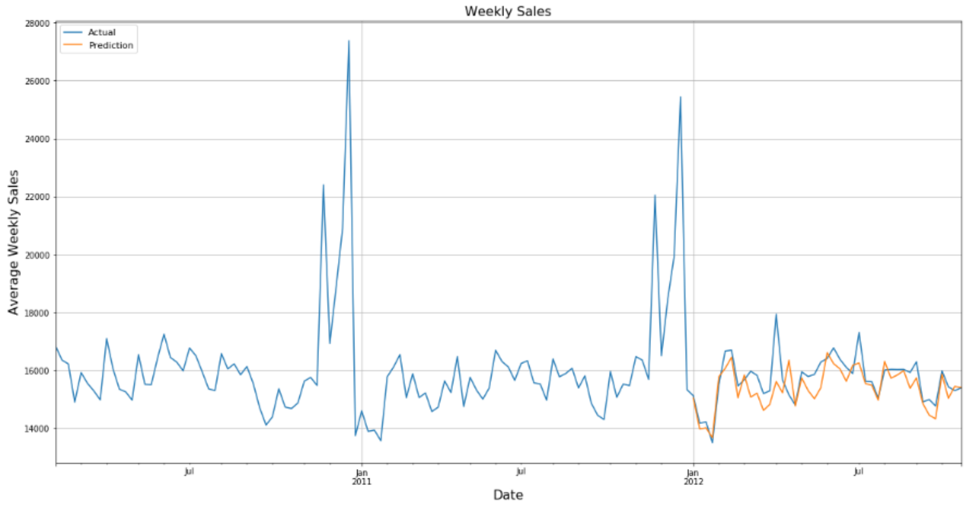


Fig.24 SARIMA compared to actual data

Holt Winter Seasonal Method:

Holt Winter seasonal method is also called Triple Exponential Smoothing algorithm as it smooth out three parameters, which are level, trend and seasonality exponentially. The beauty of the Holt Winter Seasonal package is that the preset parameters could already generate a high accuracy prediction, so far the model fit well as SARIMA. Further investigation by using performance metric would be in the comparison session.

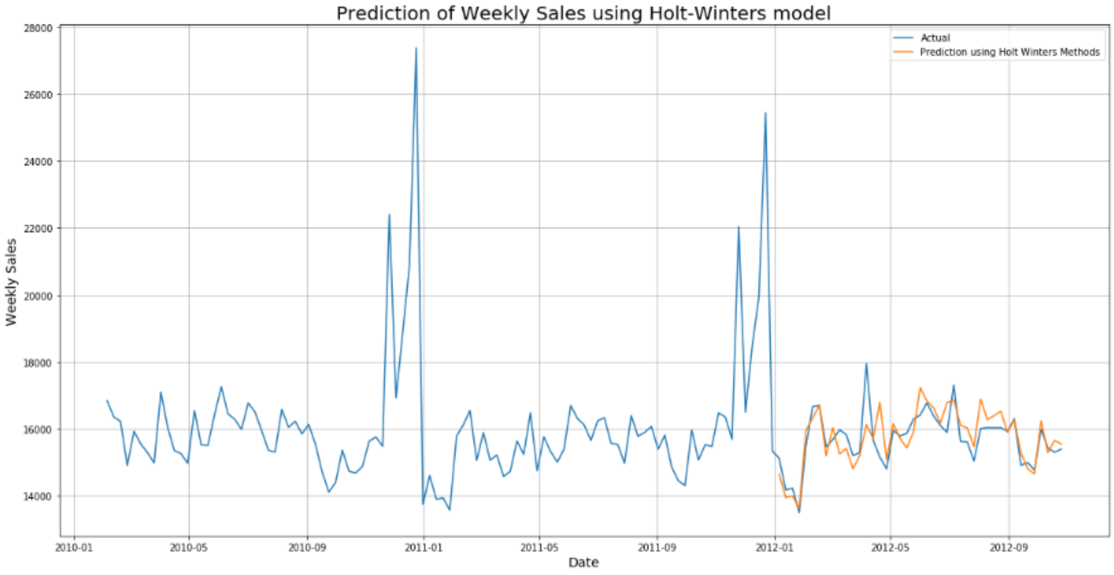


Fig.25 Holt Winter Seasonal Method compared to actual data

Regression Approach

In the previous section, only time series approaches have been used and examined as the data set consist of time stamp. In this section, machine learning regression models would be examined based on attributes relationship. The models included three common methods: Decision Tree, Random Forest Regressor, and XGBoost.

Decision Tree (or Regressor Tree)

Decision tree are supervised learning algorithm that the final output is based on comparing different values of predictors against threshold values. This technique constructs a tree model, the tree starts with the topmost node, which is the root node (class), the technique tests the data base on the root node and only two outcomes would be provided, which are the branches. The tree keeps extending until all of the data are covered in the decision tree. Decision could be implemented easily as the package could run it automatically. In the later stage of the project, parameter would be evaluating and using grid search to find the best model for decision tree.

Random Forest Regressor

Similar to Decision Tree, the random forest regressor is an ensemble regressor using many decision trees models. Instead of splitting the node base on probability, random forest regressor randomly selected subset of variables and split each node. Eventually when used to predict data, instead of based on one single decision tree, the output would be predicting base on multiple decision trees. Therefore theoretically, random forest regressor would outperform decision tree. Common parameters included number of trees, number of variables used for each split, max depth and so on. Fortunately, the model could also be automatically, further investigation on grid search would be perform in later stage of the project to find the ultimate random forest regressor model.

XGBoost

XGBoost is the short name of eXtreme Gradient Boosting, it is a custom tree building algorithm. It is also similar to decision tree, but instead, rather than training individual trees and perform predict, XGBoost train models in succession. Each successive model would be trained and improve previous error until no further improvement can be done. With decision tree, even it is being repeated, since they are trained isolated, they may end up making the same mistake.

LSTM and GRU

Note: As mentioned in the previous meeting, I have not completed this part and therefore would be skipped for now. Sorry for the inconvenience. Also, further description for each model in the regression would be included, including parameters.

Preliminary Model Comparison and Result

As below table, after preliminary comparison of the model based on MSE, MAE, RMSE, Holt Winter Seasonal Method currently shows the best score among the other models. Further investigation on the other methods and performance metrics of WMAE would also be used for comparison.

Methods\Metics	MSE	MAE	RMSE	Rank
AR	1,786,083.79	1,019.58	1,336.44	4
ARIMA	1,411,119.77	813.27	1,187.91	3
SARIMA	343,466.02	422.64	586.06	2
Holt Winter Seasonal Method	282,655.10	383.26	531.65	1
Decision Tree	21,696,159.53	1,682.75	4,657.91	7
Random Forest Regressor	12,115,351.79	1,356.02	3,480.71	6
XGBoost	10,625,660.65	1,689.27	3,259.70	5

Table 25 Time Series and Regression method comparison

References

Baccar, Y. B., ROEUFF, F., LePennec, E., d'Alché-Buc, F., Bertrand, L. A. M. Y., & Jacques, D. O. A. N. (2019). Comparative Study on Time Series Forecasting.

Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019, December). Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In *International Conference on Neural Information Processing* (pp. 462-474). Springer, Cham.

Bonnes, K. (2014). Predictive analytics for supply chains: A systematic literature review. In *21st twente student conference on IT. Netherlands*.

Goyal, A., Kumar, R., Kulkarni, S., Krishnamurthy, S., & Vartak, M. (2018). A solution to forecast demand using long short-term memory recurrent neural networks for time series forecasting. In *Midwest Decision Sciences Institute Conference*.

Hülsmann, M., Borscheid, D., Friedrich, C. M., & Reith, D. (2012). General sales forecast models for automobile markets and their analysis. *Trans. MLDM*, *5*(2), 65-86.

Jain, A., Menon, M. N., & Chandra, S. (2015). Sales Forecasting for Retail Chains.

Längkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, *42*, 11-24.

Lasek, A., Cercone, N., & Saunders, J. (2016). Restaurant sales and customer demand forecasting: Literature survey and categorization of methods. In *Smart City 360°* (pp. 479-491). Springer, Cham.

Mentzer, J. T., & Moon, M. A. (2004). *Sales forecasting management: a demand management approach*. Sage Publications.

Pai, P. F., & Lin, C. S. (2005). Using support vector machines to forecast the production values of the machinery industry in Taiwan. *The International Journal of Advanced Manufacturing Technology*, *27*(1-2), 205.

Pai, P. F., Lin, K. P., Lin, C. S., & Chang, P. T. (2010). Time series forecasting by a seasonal support vector regression model. *Expert Systems with Applications*, *37*(6), 4261-4265.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data*, *4*(1), 15.

Udom, P., & Phumchusri, N. (2014). A comparison study between time series model and ARIMA model for sales forecasting of distributor in plastic industry. *IOSR Journal of Engineering*, *4*(2), 32-38.

Wu, C. S. M., Patil, P., & Gunaseelan, S. (2018, November). Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)* (pp. 16-20). IEEE.